On Sparse, Spectral and Other Parameterizations of Binary Probabilistic Models

David Buchman[1] Mark Schmidt[2] Shakir Mohamed[1] David Poole[1] Nando de Freitas[1]

[1] Department of Computer Science, University of British Columbia, Vancouver, B.C., Canada

[2] INRIA – SIERRA Team, Laboratoire d'Informatique de l'École Normale Supérieure, Paris, France

Artificial Intelligence and Statistics (AISTATS) 2012

Introduction

Consider a general log-linear (Markov) model with binary variables $\mathbf{S} = (x_1, \ldots, x_n)$:

$$p(x) = \frac{1}{Z} \prod_{A \subseteq S} e^{\phi_A(x_A)}$$

◇ May contain potentials over arbitrary sets of variables

◇ 2^n potentials, typically most are not modeled (set to zero)

Parameterizations for each potential $\phi_A(x_A)$:

◇ "Full" / "Ising" / "Generalized Ising" / "Canonical" / "Spectral" / ...

Parameterizations – Properties

Parameterization	Complete	Minimal	Symmetric w.r.t. Values	Symmetric w.r.t. Variables	Uniquely Defined
Full	Yes	No	Yes	Yes	Yes
Ising	No	No	No	Yes	Yes
Generalized Ising	When binary	No	No	Yes	Yes
Canonical	Yes	Yes	No	No	No
Canonical (C1/C2)	Yes	Yes	No	Yes	Yes
Spectral (for binary)	Yes	Yes	Yes	Yes	Yes

Parameterizations

For Binary Variables:

Parameterization:	1-Var Pot. Bases	2-Var Pot. Bases	3-Var Potential Bases	$\phi_{ij}(x_i, x_j)$	#Params per k-Way Pot.	Total #Params
Full	$\begin{bmatrix}1\\0\end{bmatrix}$ $\begin{bmatrix}0\\1\end{bmatrix}$	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$ $\begin{bmatrix}0&1\\0&0\end{bmatrix}$ $\begin{bmatrix}0&0\\1&0\end{bmatrix}$ $\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$\begin{bmatrix}1&0\\0&0\end{bmatrix}\begin{bmatrix}0&0\\0&0\end{bmatrix}$ $\begin{bmatrix}0&1\\0&0\end{bmatrix}\begin{bmatrix}0&0\\0&0\end{bmatrix}$ ... $\begin{bmatrix}0&0\\0&0\end{bmatrix}\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$\sum_{s_1,s_2} w_{ijs_1s_2} \mathbb{I}_{x_i=s_1,x_j=s_2}$	2^k	3^n
Ising	Special treatment	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}1&0\\0&0\end{bmatrix}\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$w_{ij}\mathbb{I}_{x_i=x_j}$	1	2^n
Generalized Ising	$\begin{bmatrix}1\\0\end{bmatrix}$ $\begin{bmatrix}0\\1\end{bmatrix}$	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$ $\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$\begin{bmatrix}1&0\\0&0\end{bmatrix}\begin{bmatrix}0&0\\0&0\end{bmatrix}$ $\begin{bmatrix}0&0\\0&0\end{bmatrix}\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$\sum_{s} w_{ijs}\mathbb{I}_{x_i=x_j=s}$	2	$2^n \cdot 2$
Canonical (C1)	$\begin{bmatrix}0\\1\end{bmatrix}$	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$\begin{bmatrix}0&0\\0&0\end{bmatrix}\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$w_{ij}\mathbb{I}_{x_i\neq x_i^{ref}, x_j \neq x_j^{ref}}$	1	2^n
Spectral	$\begin{bmatrix}1\\-1\end{bmatrix}$	$\begin{bmatrix}1&-1\\-1&1\end{bmatrix}$	$\begin{bmatrix}1&-1\\-1&1\end{bmatrix}\begin{bmatrix}-1&1\\1&-1\end{bmatrix}$	$w_{ij}x_ix_j$	1	2^n

Extension to Discrete Variables With c Values:

Parameterization:	1-Var Potential Bases	2-Var Potential Bases	#Params per k-Way Pot.	Total #Params
Full	$\begin{bmatrix}1\\0\\0\end{bmatrix}$ $\begin{bmatrix}0\\1\\0\end{bmatrix}$ $\begin{bmatrix}0\\0\\1\end{bmatrix}$	$\begin{bmatrix}1&0&0\\0&0&0\\0&0&0\end{bmatrix}$ $\begin{bmatrix}0&1&0\\0&0&0\\0&0&0\end{bmatrix}$ ... $\begin{bmatrix}0&0&0\\0&0&0\\0&0&1\end{bmatrix}$	c^k	$(c+1)^n$
Ising	Special treatment	$\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$	1	2^n
Generalized Ising	$\begin{bmatrix}1\\0\\0\end{bmatrix}$ $\begin{bmatrix}0\\1\\0\end{bmatrix}$ $\begin{bmatrix}0\\0\\1\end{bmatrix}$	$\begin{bmatrix}1&0&0\\0&0&0\\0&0&0\end{bmatrix}$ $\begin{bmatrix}0&0&0\\0&1&0\\0&0&0\end{bmatrix}$ $\begin{bmatrix}0&0&0\\0&0&0\\0&0&1\end{bmatrix}$	c	$2^n c$
Canonical (C1)	$\begin{bmatrix}0\\1\\0\end{bmatrix}$ $\begin{bmatrix}0\\0\\1\end{bmatrix}$	$\begin{bmatrix}0&0&0\\0&1&0\\0&0&0\end{bmatrix}$ $\begin{bmatrix}0&0&0\\0&0&1\\0&0&0\end{bmatrix}$ $\begin{bmatrix}0&0&0\\0&0&0\\0&1&0\end{bmatrix}$ $\begin{bmatrix}0&0&0\\0&0&0\\0&0&1\end{bmatrix}$	$(c-1)^k$	c^n

Parameterizing for Learning

◇ All complete parameterizations share the same ML estimate.

◇ (parameterization, regularizer) = prior

◇ MAP: $\mathbf{w} = \arg\max_{\mathbf{w}} \left(\sum_i \log p(x_i \mid \mathbf{w}) - \mathrm{reg}(\mathbf{w}) \right)$

◇ New priors: (Spectral, *)

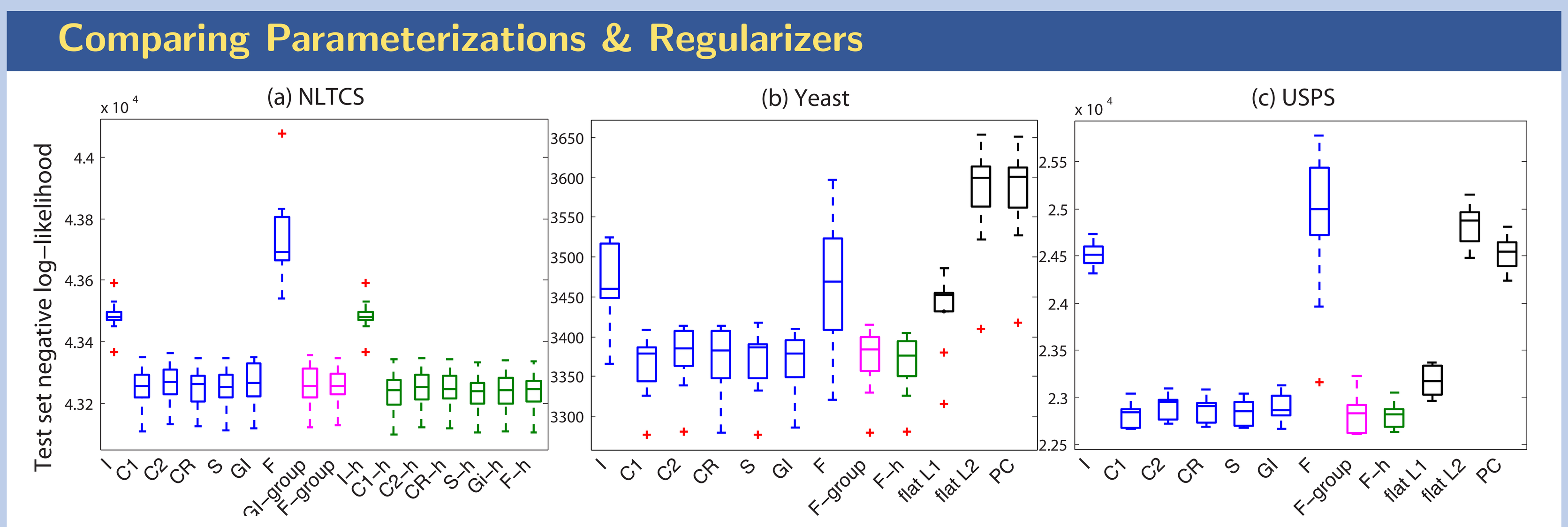
◇ What makes a good prior?

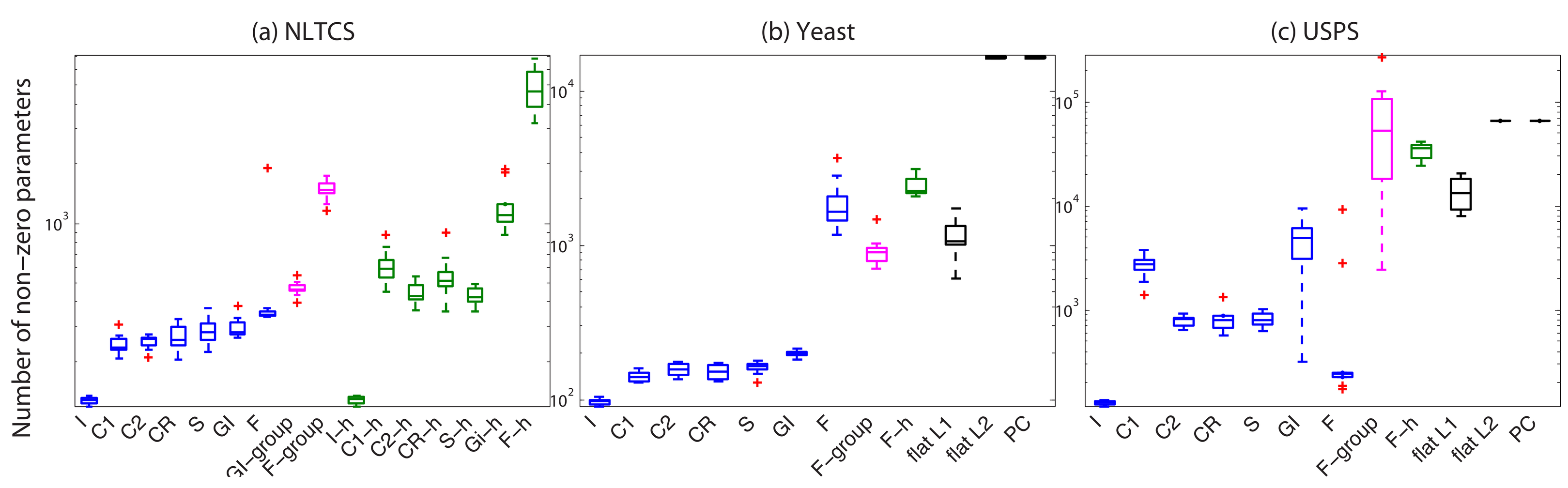
▶ Prediction accuracy

▶ Sparsity

▶ Computation

Comparing Parameterizations & Regularizers





	ℓ_1	Group ℓ_1	Hierarchical Group ℓ_1	Flat ℓ_1	Flat ℓ_2
reg(w)	$\sum_{A\subseteq S} \lambda_A \lVert w_A \rVert_1$	$\sum_{A\subseteq S} \lambda_A \lVert w_A \rVert_2$	$\sum_{A\subseteq S} \lambda_A \left(\sum_{B\supseteq A} \lVert w_B \rVert_2^2 \right)^{1/2}$	$\lambda \lVert w \rVert_1$	$\lambda \lVert w \rVert_2^2$
Parameterization	$= \sum_{A\subseteq S} \lambda_A \sum_j \lvert w_A^{(j)} \rvert$	$= \sum_{A\subseteq S} \lambda_A \left(\sum_j \left(w_A^{(j)}\right)^2 \right)^{1/2}$	$= \sum_{A\subseteq S} \lambda_A \left(\sum_{B\supseteq A} \sum_j \left(w_B^{(j)}\right)^2 \right)^{1/2}$	$= \lambda \sum_j \lvert w^{(j)} \rvert$	$= \lambda \sum_j {w^{(j)}}^2$
Ising	I	I	I-h		
Canonical – C1	C1	C1	C1-h		
Canonical – C2	C2	C2	C2-h		
Canonical – CR	CR	CR	CR-h		
Spectral	S	S	S-h	flat L1	flat L2
Generalized Ising	GI	GI-group	GI-h		
Full	F	F-group	F-h		

CR = Canonical with a random reference state.

PC = Pseudo-Counts estimate.

$\lambda_A = \mathbb{I}_{|A|\geq 2} 2^{|A|-2}\lambda$, λ chosen using cross validation.

$n \leq 16 \implies$ Partition function Z is tractable $\implies$ Can compute the exact global optimum $\implies$ Comparison results more reliable.

$\implies$ Conclusions:

◇ Use a complete & minimal parameterization (spectral / canonical).

◇ Prefer the *standard* ℓ_1.

◇ No natural reference state?

$\implies$ (Spectral, standard ℓ_1) seems the natural choice.

Spectral Interpretation

Dual representation for continuous signals:

Frequency Domain

FT (Fourier Transform) ↑ ↓ (Inverse) FT

Time Domain

Dual representation for binary distributions:

"Interactions" Domain ("Spectrum")

WHT (Walsh-Hadamard Transform) ↑ ↓ (Inverse) WHT

Log-Probability Domain

$$(\log p =) \;\; \mathbf{q} = 2^{n/2} H_n \mathbf{w}$$

$$\mathbf{w} = 2^{-n/2} H_n \mathbf{q}$$

Hadamard matrix: $H_n = 2^{-n/2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{\otimes n}$ (Kronecker power)

◇ "Spectral" seems a "natural" parameterization.

◇ "Canonical" may be "natural" when problem domain has a natural reference state.

The Statistics (Spectrum) of Binary Data Sets

◇ The spectral parameterization defines a dual "interactions" representation for binary distributions:

$$\log p(x) \iff \{w_A\} \qquad (w = \{w_A\} = \text{"interactions"})$$

◇ How can we measure the interactions in a binary *data set*?

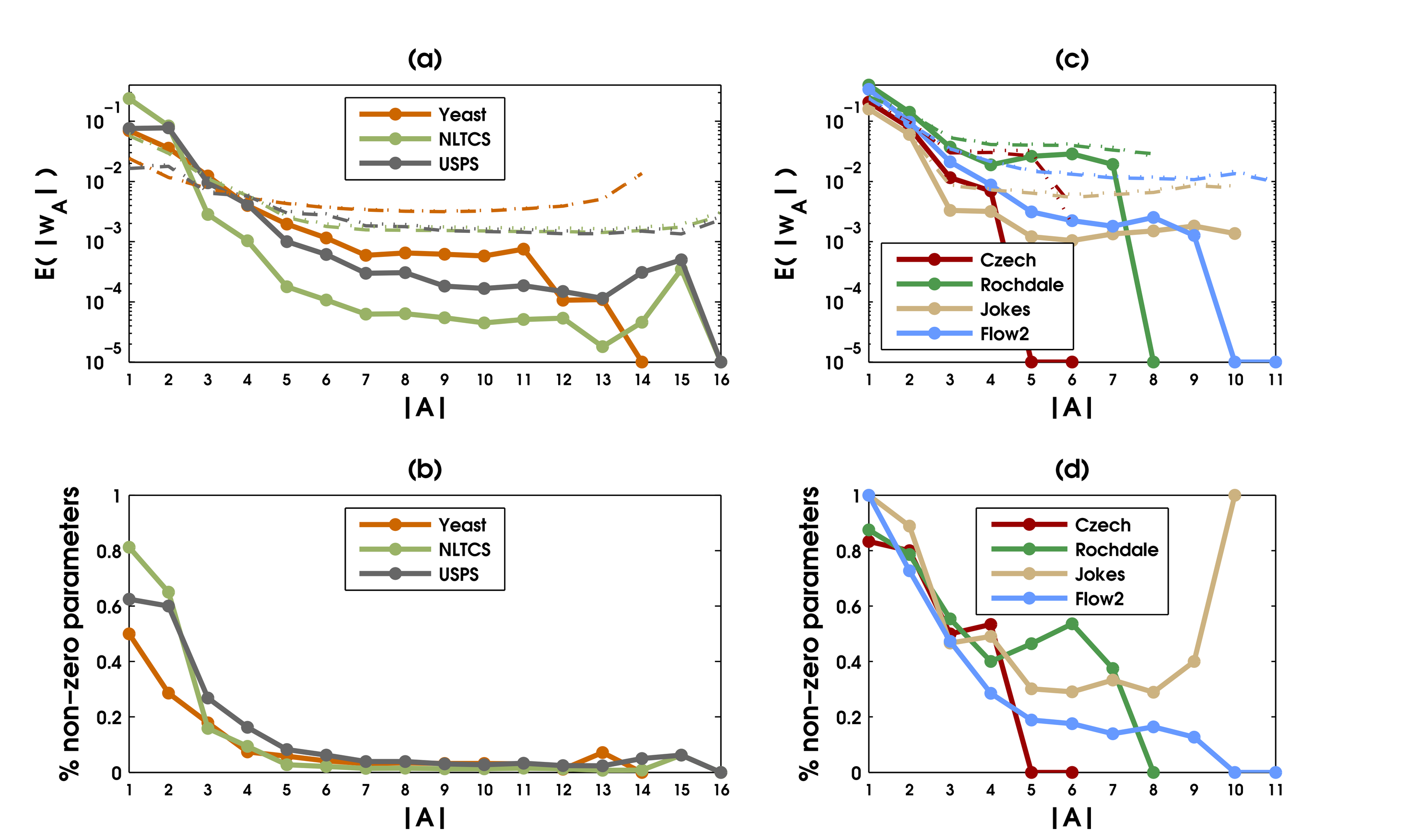
▶ Learn an (approximate) distribution $p'(x) \approx p(x)$ from the data

▶ Represent $p'(x)$ as $\{w'_A\} \approx \{w_A\}$

◇ Learning $p'(x)$:

▶ Avoid prior bias for smaller-cardinality factors: Use a parameterization which treats all $2^n - 1$ spectral parameters equally $\implies$ Use (spectral, "flat" priors) / PC / ...

▶ High modeling accuracy ($p'(x) \approx p(x)$) $\implies$ Narrow down to: (spectral, flat ℓ_1)



Solid lines = flat ℓ_1, dashed lines = flat ℓ_2, dotted lines = **PC**.

◇ Results:

▶ Higher-order interactions $\{w'_A\}$ decrease exponentially with #vars = $|A|$

▶ Confirms intuition & practical experience (adding higher potentials to models gives diminishing returns)

▶ Rationale for prior bias for lower-order factors

Select References

◇ Yvonne M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete multivariate analysis: Theory and practice.* MIT Press, 1975.

◇ Daphne Koller and Nir Friedman. *Probabilistic graphical models: Principles and techniques.* MIT Press, 2009.

◇ Mark Schmidt and Kevin P. Murphy. *Convex structure learning in log-linear models: Beyond pairwise potentials.* Artificial Intelligence and Statistics, 2010.

◇ Masashi Kato, Qian Ji Gao, Hiroshi Chigira, Hiroyuki Shindo, and Masato Inoue. *A haplotype inference method based on sparsely connected multi-body Ising model.* Journal of Physics: Conference Series, 233(1), 2010.